Machine Learning

Homework 1

Jiale Shi

January 31, 2019

1 Beta Updating from censored likelihood and MLE for uniform distribution

A Suppose we toss a coin $n = 5$ times. Let X be the number of heads. Assume that we observe that there are fewer than 3 heads, however, we don't known the exact number. We consider the prior probability of heads to be $p(\theta) = \text{Beta}(\theta|1, 1)$. Compute the posterior $p(\theta|X < 3)$ up to normalization constant (i.e., derive expression proportional to $p(\theta|X < 3)$).

Solution:

Since the prior probability of heads is $p(\theta) = \text{Beta}(\theta|1, 1)$, it is equal to uniform distribution and independent with θ.

$$\begin{aligned} p(\theta|X < 3) &\propto p(X < 3|\theta)p(\theta) \\ &\propto p(X < 3|\theta) \\ &= p(X = 0|\theta)p(\theta) + p(X = 1|\theta)p(\theta) + p(X = 2|\theta)p(\theta) \\ &= 1 \cdot \theta^0(1-\theta)^5 + 5 \cdot \theta^1(1-\theta)^4 + 10 \cdot \theta^2(1-\theta)^3 \\ &= \text{Bin}(0|\theta, 5) + \text{Bin}(1|\theta, 5) + \text{Bin}(2|\theta, 5) \end{aligned} \tag{1}$$

B Consider a uniform distribution centred on 0 with width $2a$. The density function is given as

$$p(x) = \frac{1}{2a}\mathbb{I}(x \in [-a, a]) \tag{2}$$

(a) Given a dataset $x_1, ..., x_n$,what is the MLE of a?

Solution:

$$\begin{aligned} p(D|a) &= \left[\frac{1}{\text{size}(a)}\right]^n = \left[\frac{1}{2a}\right]^n \\ \hat{a}^{mle} &= \text{argmax} p(D|a) = \max \text{abs}(x_1, ..., x_n) \end{aligned} \tag{3}$$

The MLE of a is the maximum absolute value of the given data set $x_1, ..., x_n$.

(b) What probability would the model assign to $\hat{x}_{n+1}$ using the MLE estimate of a?

Solution:

$$p(\hat{x}_{n+1}|\hat{a}^{mle}) = \left\{ \begin{array}{rll} \frac{1}{2\hat{a}^{mle}} & \text{for} & |\hat{x}_{n+1}| \leq \hat{a}^{mle} \\ 0 & \text{for} & |\hat{x}_{n+1}| > \hat{a}^{mle} \end{array} \right\} \tag{4}$$

(c)Do you see any problem with the above approach? If yes, briefly suggest a better alternative.

Solution:

In (b),

$$p(\hat{x}_{n+1}|\hat{a}^{mle}) = 0 \quad \text{for} |\hat{x}_{n+1}| > \hat{a}^{mle} \tag{5}$$

Therefore, using MLE can perform quite poorly when the sample data set size is small. This is called the **zero count problem** or the **sparse data problem**,

and frequently occurs when estimating counts from small amounts of data. A better alternative would be calculate the posterior by using Bayesian analysis and choosing a suitable prior.

2 Bayesian analysis for Uniform distribution and the Taxicab problem

A Consider the uniform distribution $\mathcal{U}(0,\theta)$. Given the Pareto prior, the joint distribution of θ and $\mathcal{D} = (x_1, ..., x_N)$ is

$$p(\mathcal{D},\theta) = \frac{Kb^K}{\theta^{N+K+1}}\mathbb{I}(\theta \geq \max(\mathcal{D})) \tag{6}$$

Let $m = \max(\mathcal{D})$. The evidence (the probability that all N samples came from the same uniform distribution) is

$$\begin{aligned} p(\mathcal{D}) &= \int_m^\infty \frac{Kb^K}{\theta^{N+K+1}}d\theta \\ &= \left\{ \begin{array}{lll} \frac{K}{(N+K)b^N} & \text{if} & m \leq b \\ \frac{Kb^K}{(N+K)m^{N+K}} & \text{if} & m > b \end{array} \right\} \end{aligned} \tag{7}$$

Derive the posterior $p(\theta|\mathcal{D})$ and show that it can be expressed as a Pareto distribution.
Solution:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D},\theta)}{p(\mathcal{D})} = \frac{Kb^K}{\theta^{N+K+1}}\mathbb{I}(\theta \geq \max(\mathcal{D},b))\frac{1}{p(\mathcal{D})} \tag{8}$$

The first case, if $m \leq b$,

$$\begin{aligned} p(\theta|\mathcal{D}) &= \frac{Kb^K}{\theta^{N+K+1}}\mathbb{I}(\theta \geq \max(\mathcal{D},b))\frac{(N+K)b^N}{K} \\ &= \frac{(N+K)b^{N+K}}{\theta^{N+K+1}}\mathbb{I}(\theta \geq b)) \\ &= Pareto(\theta|N+K,b) \end{aligned} \tag{9}$$

The second case, if $m > b$,

$$\begin{aligned} p(\theta|\mathcal{D}) &= \frac{Kb^K}{\theta^{N+K+1}}\mathbb{I}(\theta \geq \max(\mathcal{D},b))\frac{Kb^K}{(N+K)m^{N+K}} \\ &= \frac{(N+K)m^{N+K}}{\theta^{N+K+1}}\mathbb{I}(\theta \geq m)) \\ &= Pareto(\theta|N+K,m) \end{aligned} \tag{10}$$

Therefore,

$$p(\theta|\mathcal{D}) = Pareto(\theta|N+K,\max(m,b)) \tag{11}$$

B Suppose you arrive in a new city and see a taxi numbered 100. How many taxis are there in this city? Let us assume taxis are numbered sequentially

as integers starting from 0, up to some unknown upper bound θ. Hence the likelihood function is $p(x) = \mathcal{U}(0, \theta)$.
(a) Suppose we see one taxi numbered 100. Using an non-informative prior on θ of the form $p(\theta) = \text{Pa}(\theta|0, 0) \propto \frac{1}{\theta}$. What is the posterior $p(\theta|\mathcal{D})$?
Solution:
$p(\theta) = \text{Pa}(\theta|0, 0) \propto \frac{1}{\theta}$, therefore, $b = 0$, $K = 0$.
Suppose we see one taxi numbered 100, therefore, $\mathcal{D} = 100$, $m = 100$, $N = 1$. Using the Bayesian analysis result from Part. A.

$$\begin{aligned} p(\theta|\mathcal{D}) &= Pareto(\theta|N + K, \max(m, b)) \\ &= Pareto(\theta|1 + 0, \max(100, 0)) \\ &= Pareto(\theta|1, 100) \end{aligned} \tag{12}$$

(b) Compute the posterior mean, mode and median number of taxis in the city, if such quantities exist.
Solution:
From wikipedia,
The posterior mean of $Pareto(\theta|\alpha, x_m)$ is equal to $E[\theta|D] = \frac{\alpha x_m}{\alpha - 1}, \alpha > 1$. But in this Pareto distribution,$Pareto(\theta|1, 100)$, $\alpha = 1$ and $x_m = 100$. Therefore, the posterior mean does not exist.
The mode of this Pareto distribution, mode$(\theta|D) = x_m = 100$
The median of this Pareto distribution, median$(\theta|D) = x_m(2)^{\frac{1}{\alpha}} = 100(2)^1 = 200$
(c) Rather than trying to compute a point estimate of the the number of taxis, we can compute the predictive density over the next taxicab number using

$$p(D^{'}|\mathcal{D}, \alpha) = \int p(D^{'}|\theta)p(\theta|\mathcal{D}, \alpha)d\theta = p(D^{'}|\beta) \tag{13}$$

where $\alpha = (b, K)$ are the hyperparameters and $\beta = (c, N + K)$ are the updated hyperparameters. Now consider the case $\mathcal{D} = \{m\}$, and $D^{'} = \{x\}$. Using Part A of this problem, write down the expression for $p(x|\mathcal{D}, \alpha)$. As above, consider, $b = K = 0$.
Solution:
In order to compute the predictive density over the next taxicab number m, we need to calculate the posterior $p(\theta|\mathcal{D})$ with $N = 1$,$b = K = 0$

$$\begin{aligned} p(\theta|\mathcal{D}) &= Pareto(\theta|N + K, \max(m, b)) \\ &= Pareto(\theta|1 + 0, \max(m, 0)) \\ &= Pareto(\theta|1, m) \end{aligned} \tag{14}$$

We can use this posterior on $\mathcal{D}$ as a "prior" for the inference on the next taxicab $D^{'}$. This new "prior" for the next taxicab $p(\theta) = Pareto(\theta, K^{'} = 1, b^{'} = m)$.$N^{'} = 1$ and $m^{'} = \max(D^{'}) = x$.

Using equation from part A.

$$
\begin{aligned}
p(x|\mathcal{D},\alpha) &= \frac{K^{'}}{(N^{'}+K^{'}){b^{'}}^{N^{'}}}\mathbb{I}(m^{'} \leq b^{'}) + \frac{K^{'}{b^{'}}^{K^{'}}}{(N^{'}+K^{'}){m^{'}}^{N^{'}+K^{'}}}\mathbb{I}(m^{'} > b^{'}) \\
&= \frac{1}{2m}\mathbb{I}(x \leq m) + \frac{m}{2x^2}\mathbb{I}(x > m)
\end{aligned} \tag{15}
$$

(d) Use the predictive density formula to compute the probability that the next taxi you will see (say, the next day) has the number 100, 50 or 150.
Solution:

$$
\begin{aligned}
p(x=100|\mathcal{D},\alpha) &= \frac{1}{2m}\mathbb{I}(100 \leq m) + \frac{m}{20000}\mathbb{I}(100 > m) \\
p(x=50|\mathcal{D},\alpha) &= \frac{1}{2m}\mathbb{I}(50 \leq m) + \frac{m}{5000}\mathbb{I}(50 > m) \\
p(x=150|\mathcal{D},\alpha) &= \frac{1}{2m}\mathbb{I}(150 \leq m) + \frac{m}{45000}\mathbb{I}(150 > m)
\end{aligned} \tag{16}
$$

(e) Briefly describe (1-2 sentences) some ways you might make the model more accurate at prediction.
Solution:
(1) Use an informative prior on θ instead of non-informative prior.
(2) Use a discrete probability distribution since m and b are discrete integer in this problem. In addition, instead of using integral, use summation.

3 Naive Bayes

3.1

Consider a Naive Bayes model (multivariate Bernouli version) for spam classification with the vocabulary V ="secret", "offer", "low", "price", "valued", "customer", "today", "dollar", "million", "sports", "is", "for", "play", "healthy", "pizza". We have the following example, spam messages "million dollar offer", "secret offer today","secret is secret" and normal messages, "low price for valued customer","play secret sports today","sports is healthy","low price pizza". Given the MLEs for the following parameters: θ_{spam}, $\theta_{\text{secret|spam}}$, $\theta_{\text{secret|non-spam}}$, $\theta_{\text{sports|non-spam}}$, $\theta_{\text{dollar|spam}}$.

Solution: MLE

$$\hat{\theta}_k^{mle} = \frac{N_k}{N} \tag{17}$$

$$\begin{aligned} N_{\text{spam}} &= 3 \\ N_{\text{non-spam}} &= 4 \\ N_{\text{secret,spam}} &= 2 \\ N_{\text{secret,non-spam}} &= 1 \\ N_{\text{sports,non-spam}} &= 2 \\ N_{\text{dollar|spam}} &= 1 \end{aligned} \tag{18}$$

$$\begin{aligned} \theta_{\text{spam}} &= \frac{3}{3+4} = \frac{3}{7} \\ \theta_{\text{secret|spam}} &= \frac{2}{3} \\ \theta_{\text{secret|non-spam}} &= \frac{1}{4} \\ \theta_{\text{sports|non-spam}} &= \frac{2}{4} = \frac{1}{2} \\ \theta_{\text{dollar|spam}} &= \frac{1}{3} \end{aligned} \tag{19}$$

3.2

Write a code for a Naive Bayes classifier (both training and prediction) and use it for the two data sets provided at this link. Both the data sets have four variables namely, Xtrain, Xtest, ytrain and ytest. Train the Naive Bayes classifier based on Xtrain and ytrain and test it based on Xtest and ytest. Report the misclassification rate for both the data sets.

Solution:

For first dataset "Problem3b1.mat", there are 4 classes. Misclassification rate on train is 18.92%, on test is 19.3%.

For second dataset "Problem3b2.mat", there are 2 classes. Misclassification rate on train is 8.33%, on test is 18.67%.

4 Emperical Bayes

Consider the problem of predicting cancer rates in various cities. In particular, suppose we measure the number of people in various cities, N_i, and the number of people who died of cancer in those cities, x_i. We assume $x_i \sim \mathrm{Bin}(\mathrm{N_i}, \theta_\mathrm{i})$, and we want to estimate the cancer rates θ_i. One way is to estimate the parameters separately, but this will inevitably suffer from the sparse data problem. On the other hand, parameter trying for this problem will be too strong an assumption. A compromise is to assume that the parameters θ_i are similar but there may be city-specific variations. This can be modelled by assuming that the θ_i are drawn from some common distribution. Assuming that $\theta_i \sim \mathrm{Bin(a, b)}$, the joint distribution can be written as:

$$p(\mathcal{D}, \theta, \eta|N) = p(\eta) \prod_{i=1}^{N} \mathrm{Bin}(\mathrm{x_i}|\theta_\mathrm{i}, \mathrm{N_i})\mathrm{Beta}(\theta_\mathrm{i}|\eta) \tag{20}$$

where $\eta = (a, b)$. With this set up, the problem can be solved using two methods. In the first method, we can adopt a fully Bayesian approach by assigning an appropriate prior. However, with this setup, we will have to resort to some sampling based approach which will be covered later in the course. The other method is to estimate point estimates for η by using $\mathcal{D}$. This method is known as the Empirical Bayes. Note that this method violates the basic assumption of Bayesian statistics as the prior depends on the data.

Write a computer code for estimating the parameters, θ_i by using empirical Bayes. For estimating the parameters η is empirical Bayes, you can utilize the fixed point method proposed by Minka. The MATLAB code for the fixed point method and the cancer data set can be found at this link. The data set has a structure data which, in turn, has two variables y (No. of persons with cancer) and n (Total population of the city).

Solution:

We use Empirical Bayes to deal with this problem. The reasons why we use Empirical Bayes have been partially stated in the problem. I would state again: if we estimate them separately, this would suffer from the sparse data problem (underestimation of the rate of cancer due to small N_i); if we assume all the θ_i are the same, this is called parameter tying, and this assumption that all the cities have the same rate is a rather strong one. A compromise approach is to assume that the θ_i are similar, but that there may be city-specific variations. The data-poor cities borrow statistical strength from the data-rich city.

Note that it is important that we get $\eta = (a, b)$ from the data. If we just assume it to be constant, the θ_i will be conditionally independent for all the cities, and there will be no information flow between them. By contrast, by treating $\eta = (a, b)$ as unknown (hidden variable), we allow the data-poor cities to borrow statistical strength from the data-rich ones.

If we just use MLE (in third row of Figure 1), we would find that the MLE of city 20 is 0 since city 20 has a very smaller population. However, it is not reliable because this would suffer from the sparse data problem. We compute the joint

posterior $p(\eta, \theta|\mathcal{D})$. From this we can get the posterior marginals $p(\theta_i|\mathcal{D})$ and posterior means $\mathbb{E}(a/a+b|\mathcal{D})$ (in fourth row of Figure 1). The posterior mean is shrunk towards the pooled estimate more strongly for cities with small sample sizes, like city 20. If we use posterior mean, its rate is shrunk more towards to the pop mean. Even for city 10 and 19, their posterior mean are also shrunk towards to the pop mean.

Figure 2 shows the 95% posterior credible intervals for θ_i for all cities. City 15 has a very large population and has small posterior uncertainty. This data has the largest impact on the posterior estimate of η, which would in turn influence the estimate of the cancer rates for other cities.

Figure 1 and Figure 2 are created by using the matlab code offered and adding some modifications. The analyse is based on Zabaras's lectures and Murphy's book *Machine Learning*.

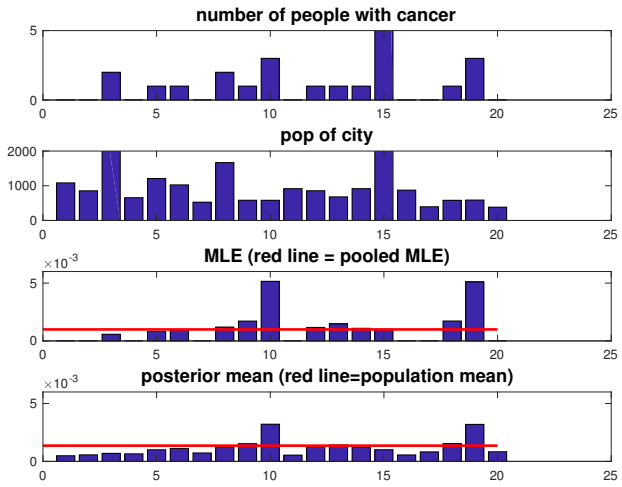

Figure 1: Results of fitting the model

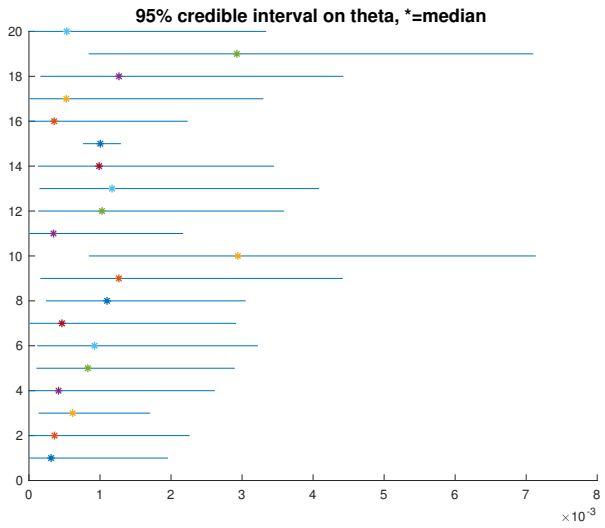

Figure 2: Posterior 95% credible intervals on the cancer rates.

5 Reject option in classifiers and Newsvendor problem

A In many classification problems one has the option either of assigning x to class j or, if you are too uncertain, of choosing the reject option. If the cost for rejects is less than the cost of falsely classifying the object, it may be the optimal action. Let α_i mean you choose action i, for $i =, ..., C+1$ where C is the number of classes and $C+1$ is the rejection option. Let $Y = j$ be the true (but unknown) state of nature. We define the loss function as

$$\lambda(\alpha_i|Y=j) = \left\{ \begin{array}{lll} 0 & \text{if} & i = j \text{and} i, j \in 1, ..., C \\ \lambda_r & \text{if} & i = C+1 \\ \lambda_s & \text{otherwise} & \end{array} \right\} \tag{21}$$

In other words you incur 0 loss if you correctly classify, you incur λ_r loss if you choose the reject option and you incur λ_s loss if you make a substitution error (misclassification).
(a) Show that the minimum risk is obtained at $Y = j$ if

$$p(Y=j|x) \geq p(Y=k|x) \forall k \tag{22}$$

and

$$p(Y=j|x) \geq 1 - \frac{\lambda_r}{\lambda_s} \tag{23}$$

otherwise we decide to reject.
Solution:
If we choose to assign a class $\alpha_k \in (\alpha_1, ..., \alpha_C) \forall k$

$$\rho(a|x) = \sum_{i=1}^{C} p(Y = i|x) L(Y = k, \alpha_k) = \sum_{i \neq k} p(Y = i|x)\lambda_s = (1 - p(Y = k|x))\lambda_s \tag{24}$$

If we choose the reject option.

$$\rho(a|x) = \sum_{i=1}^{C} p(Y = i|x) L(Y = i, \alpha_k) = \sum_{i=1}^{C} p(Y = i|x)\lambda_r = \lambda_r \tag{25}$$

The risk of obtaining at $Y = j$ is: $(1 - p(Y = j|x))\lambda_s$

If the minimum risk is obtained at $Y = j$. Therefore, the risk of obtaining at $Y = j$ should be smaller than both the risks obtained before.

$$\begin{aligned} &(1 - p(Y = j|x))\lambda_s \leq (1 - p(Y = k|x))\lambda_s \\ &p(Y = j|x) \geq p(Y = k|x) \end{aligned} \tag{26}$$

and

$$\begin{aligned} &(1 - p(Y = j|x))\lambda_s \leq \lambda_r \\ &p(Y = j|x) \geq 1 - \frac{\lambda_r}{\lambda_s} \end{aligned} \tag{27}$$

(b) Describe Qualitatively what happens as $\frac{\lambda_r}{\lambda_s}$ is increased from 0 to 1.
Solution:
If $\frac{\lambda_r}{\lambda_s} = 0$, therefore $\lambda_r = 0$ and there is no risk in choose the reject option. Then consider the (a), in order to satisfy that the minimum risk is obtained at $Y = j$, $p(Y = j|x) \geq 1 - \frac{\lambda_r}{\lambda_s} = 1$, therefore, $p(Y = j|x) = 1$.
If $\frac{\lambda_r}{\lambda_s} = 1$, $\lambda_r = \lambda_s$. This means that the loss of choosing the reject option and making a substitution error (misclassification) are the same. Then the reject option is no valuable. It is better to choose to assign a class since there are chances to meet the true class.
If $0 < \frac{\lambda_r}{\lambda_s} < 1$, with $\frac{\lambda_r}{\lambda_s}$ increases, choosing a reject option becomes less and less interesting and valuable since the loss of reject option becomes larger and larger and become closer and closer to the loss of misclassification.
B Suppose you are trying to decide how much quantity Q of some product (e.q., newspapers) to buy to maximize your profits. The optimal amount will depend on how much demand D you think there is for your product, as well as its cost to you C and its selling price P. Suppose D is known but has pdf $f(D)$ and cdf $F(D)$. We can evaluate the expected profit by considering two cases: if $D > Q$, then we sell all Q items, and make profit $\pi = (P - C)Q$; but if $D < Q$, we only sell D items, at profit $(P - C)D$, but have wasted $C(Q - D)$ on the unsold

items. So we expected profit if we by quantity Q is

$$\mathbb{E}_\pi = \int_Q^\infty (P-C)Qf(D)dD + \int_0^Q (P-C)Df(D)dD - \int_0^Q C(Q-D)f(D)dD \tag{28}$$

Show that the optimal quantity Q^* (which maximizes the expected profit) satisfies

$$F(Q^*) = \frac{P-C}{P} \tag{29}$$

Solution:

$$\begin{aligned}
\mathbb{E}_\pi &= \int_Q^\infty (P-C)Qf(D)dD + \int_0^Q (P-C)Df(D)dD - \int_0^Q C(Q-D)f(D)dD \\
&= (P-C)Q\int_Q^\infty f(D)dD + (P-C)\int_0^Q Df(D)dD - CQ\int_0^Q f(D)dD + C\int_0^Q Df(D)dD \\
&= (P-C)Q(1-F(Q)) + P\int_0^Q Df(D)dD - CQF(Q) \\
&= (P-C)Q - PQF(Q) + \int_0^Q Df(D)dD
\end{aligned} \tag{30}$$

The optimal quantity Q^* which maximizes the expected profit, means $\frac{d\mathbb{E}_\pi}{dQ}|_{Q=Q^*} = 0$

$$\begin{aligned}
&\frac{d\mathbb{E}_\pi}{dQ} = (P-C) - PF(Q) - PQf(Q) + PQf(Q) = 0 \\
&F(Q^*) = \frac{P-C}{P}
\end{aligned} \tag{31}$$